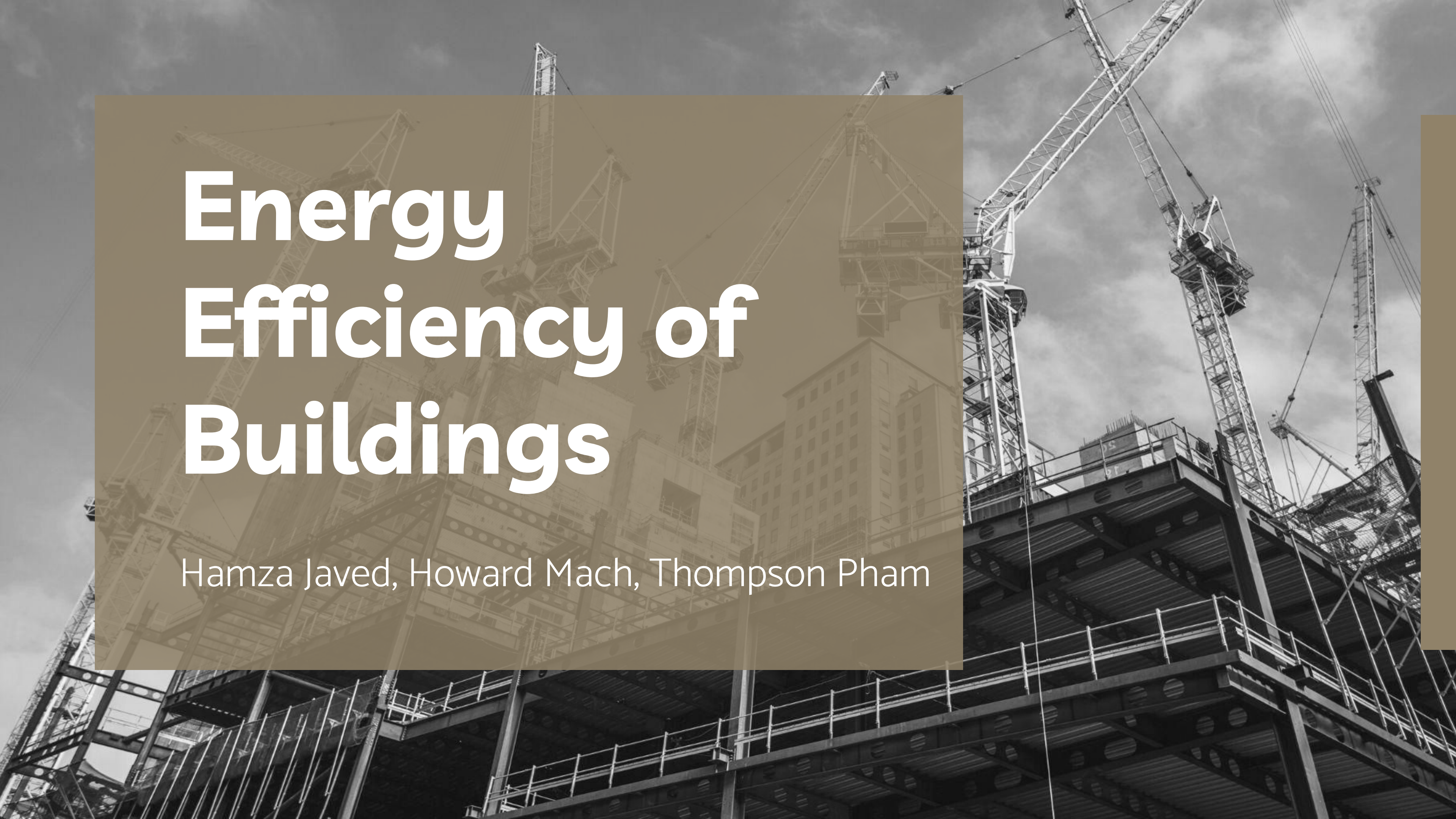
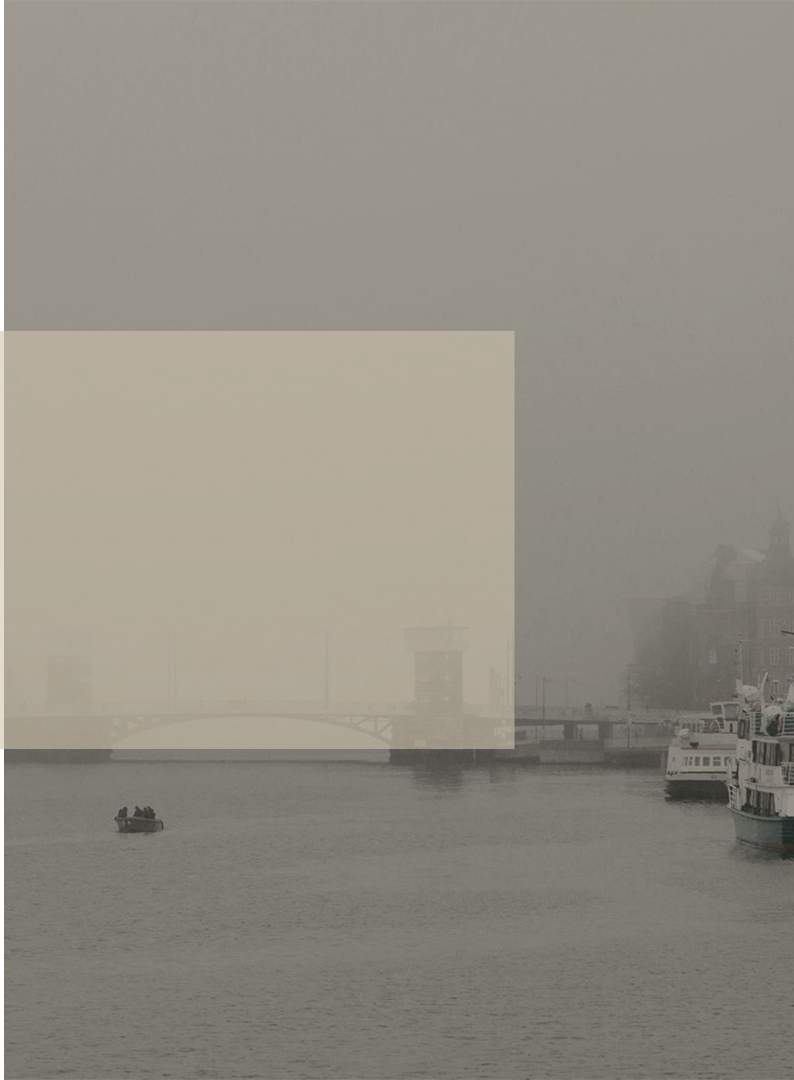


Energy Efficiency of Buildings

Hamza Javed, Howard Mach, Thompson Pham

OBJECTIVE

We plan to use principal component analysis to reduce the number of variables into a smaller set that captures the most significance. From there, we will use principal component regression to make our predictions and find our final model.



Background

Japan

In 2022 , Japan experienced a massive heat wave resulting in the government asking households and businesses to turn off lights to sustain air conditioning.

Britain

In comparison to other first world countries, buildings in Britain are generally much older. As a result, they have worse insulation temperatures..

Dataset Information

Sources

UC Irvine Machine Learning Repository
By Athanasios Tsanas
& Angeliki Xifara

Goal

Find the most influential variables for our target variable.

Size

768 observations and **8** features. Approximately 192 unique building shapes with 4 different orientations which gives us 768 unique observations

Predictors

- Relative Compactness: Real Number
- Surface Area (m2): Real Number
- Wall Area (m2): Real Number
- Roof Area (m2): Real Number
- Overall Height (m): Real Number
- Orientation (2:North, 3:East, 4:South, 5:West): Integer
- Glazing Area (% of wall is glass): Real Number
- Glazing Area Distribution (glass windows/doors): Integer

Target

Heating Load (kWh/m^2): Real Number -
Cooling Load (kWh/m^2): Real Number -

Data Cleaning & Normalization

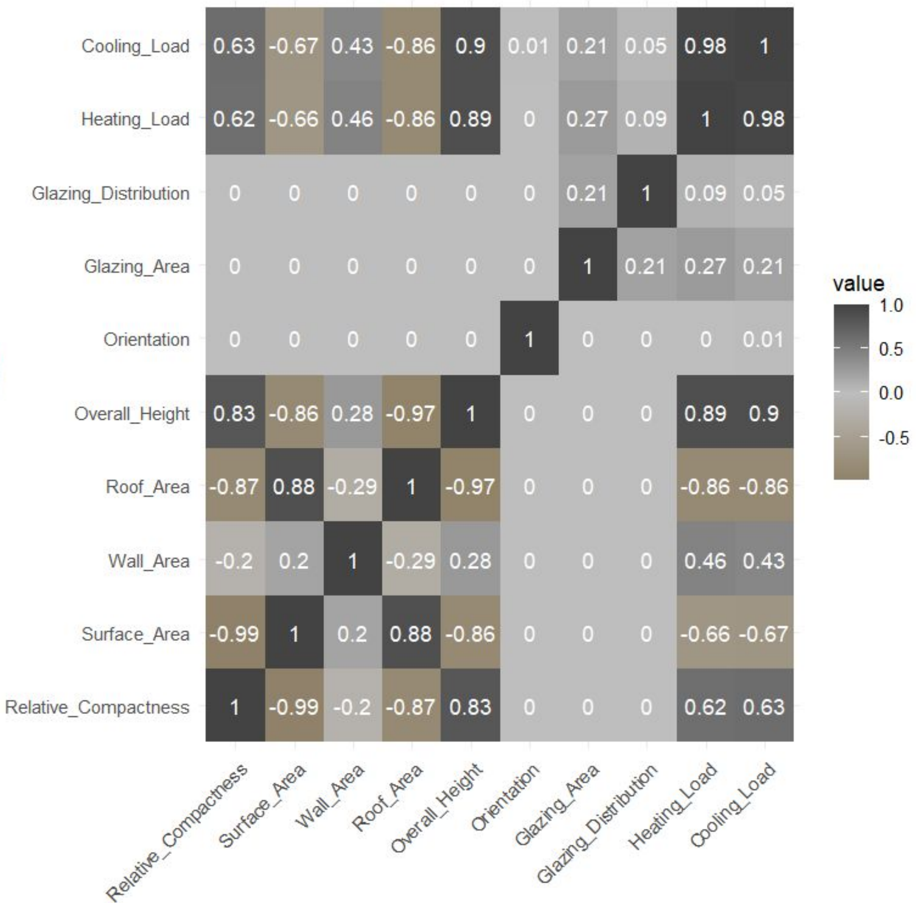
Missing Values : None

We felt all of presented variables in the dataset aligned with our goal, so we kept did not remove any.

Note: Dataset is machine generated

Our predictors have different magnitudes and ranges, so we applied a scale function to our dataset to normalize it.

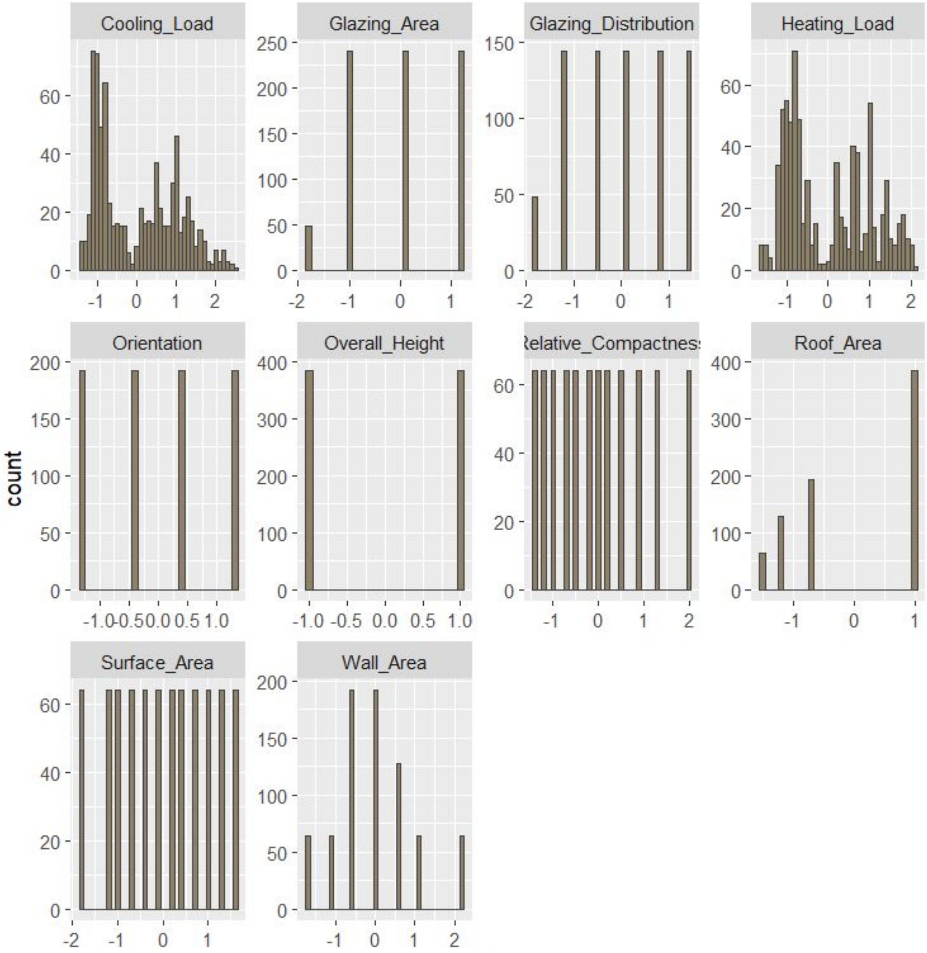
Exploratory Analysis - Correlation



	Relative_Compactness	Surface_Area	Wall_Area	Roof_Area	Overall_Height	Orientation	Glazing_Area	Glazing_Distribution	Heating_Load	Cooling_Load
Cooling_Load	0.63	-0.67	0.43	-0.86	0.9	0.01	0.21	0.05	0.98	1
Heating_Load	0.62	-0.66	0.46	-0.86	0.89	0	0.27	0.09	1	0.98
Glazing_Distribution	0	0	0	0	0	0	0.21	1	0.09	0.05
Glazing_Area	0	0	0	0	0	0	1	0.21	0.27	0.21
Orientation	0	0	0	0	0	1	0	0	0	0.01
Overall_Height	0.83	-0.86	0.28	-0.97	1	0	0	0	0.89	0.9
Roof_Area	-0.87	0.88	-0.29	1	-0.97	0	0	0	-0.86	-0.86
Wall_Area	-0.2	0.2	1	-0.29	0.28	0	0	0	0.46	0.43
Surface_Area	-0.99	1	0.2	0.88	-0.86	0	0	0	-0.66	-0.67
Relative_Compactness	1	-0.99	-0.2	-0.87	0.83	0	0	0	0.62	0.63

From the correlation matrix plot, we can see that many predictors are **linearly independent** from each other.

However, there are some predictors that are **linearly dependent** of each other. We will investigate this later.

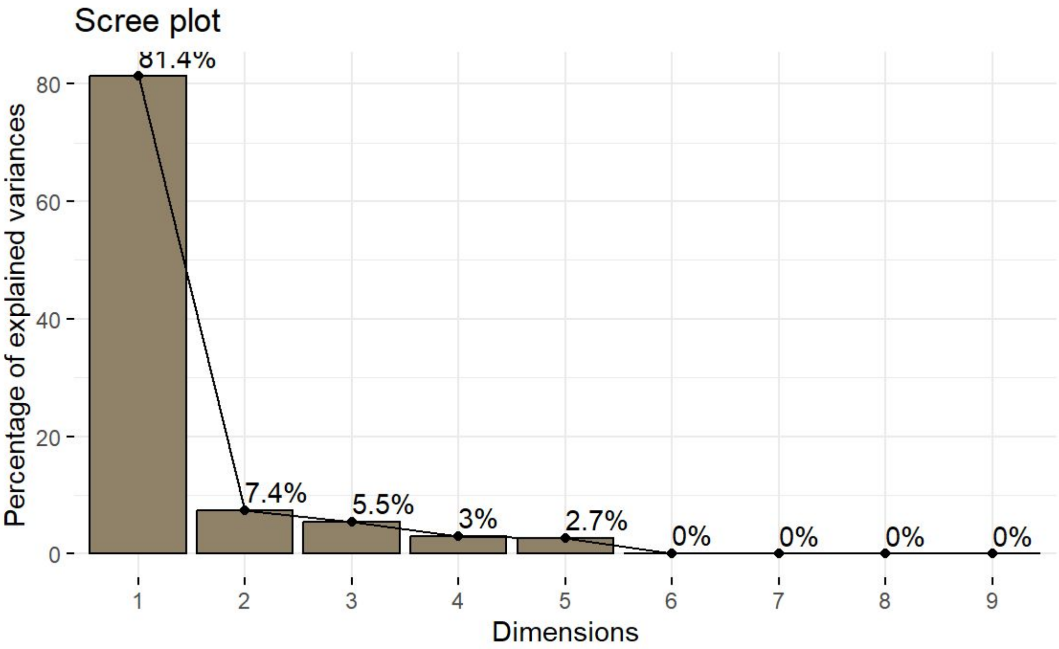


Exploratory Analysis - Variable Distribution

Since our data is simulated, many of the variable distributions are uniform.

Our scale function does not alter the content of our data, so should expect the spread of the data to be the same before and after.

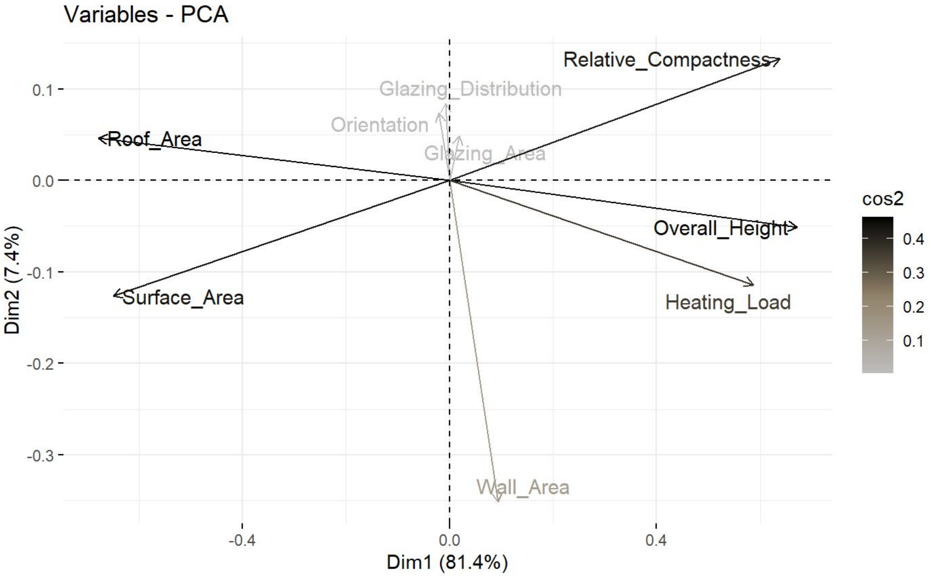
Principal Component Analysis (PCA)



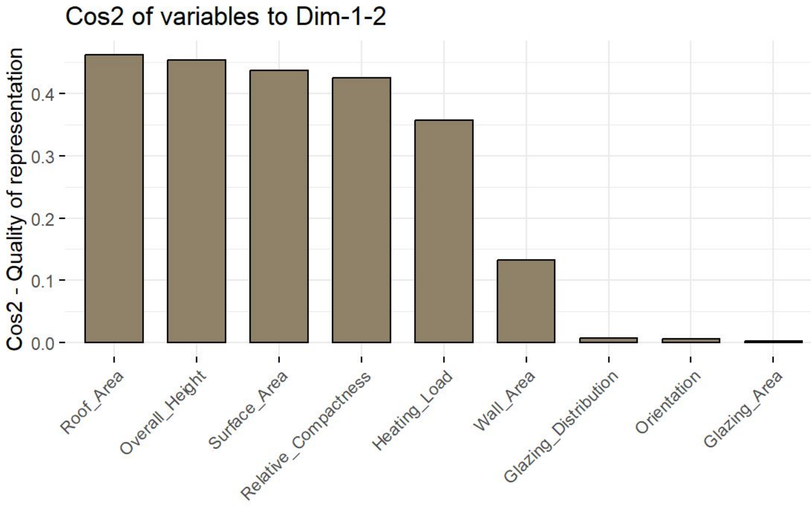
We utilized PCA to reduce the dimensionality of our data. From the the summary above we can see that component 1 and component 2 explain about 90% of our data's variance.

Therefore, it is sufficient to say that the first 2 principle components accurately represent our data.

Principal Component Analysis (PCA)



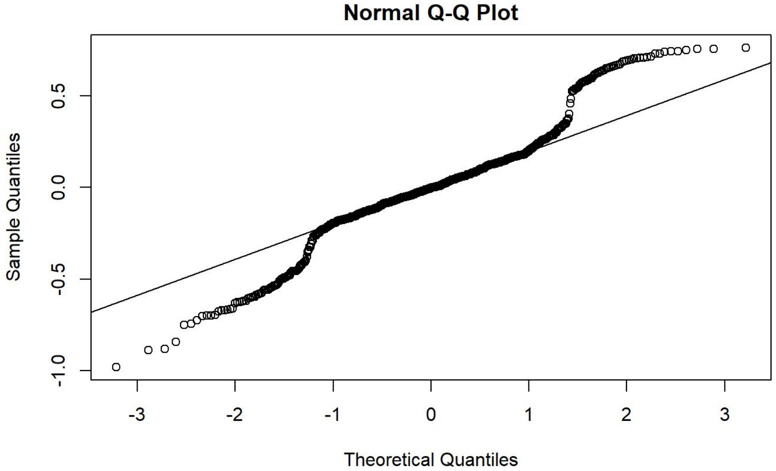
The following graphs visualize the importance of each variable on the first two principal components. We can see that **Roof Area**, **Overall Height**, **Surface Area**, and **Relative Compactness** are the most influential predictors.

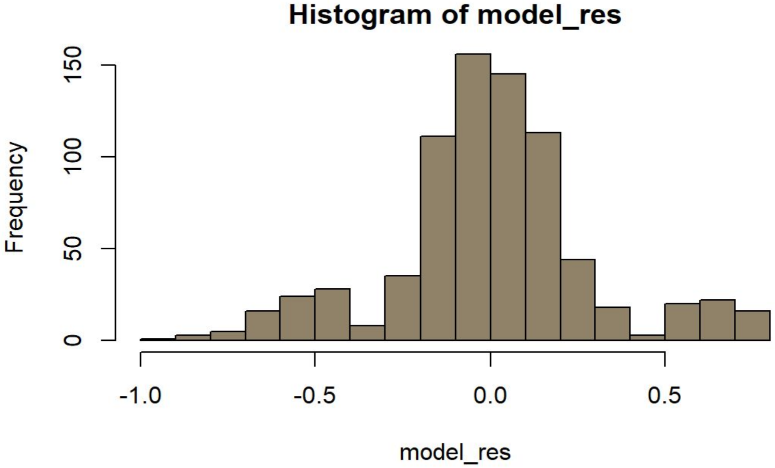


Multiple Linear Regression

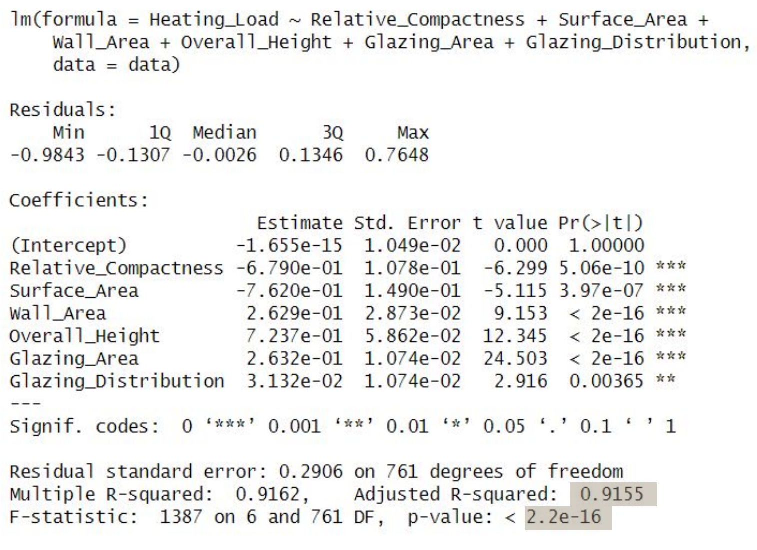
Before we can apply a linear regression model, we have to check certain assumptions. We need to check weather the residuals are normally distributed and linear.

From the Q-Q plot we can conclude that the residuals are mostly linear. The histogram indicates that the residuals are mostly normally distributed. We have checked our assumptions, so we can move on to regression.





ANOVA

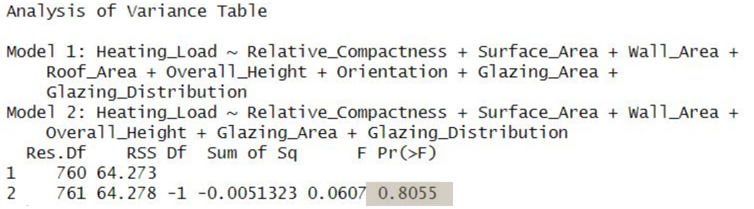


```
lm(formula = Heating_Load ~ Relative_Compactness + Surface_Area +
    Wall_Area + Overall_Height + Glazing_Area + Glazing_Distribution,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9843 -0.1307 -0.0026  0.1346  0.7648

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.655e-15  1.049e-02   0.000  1.00000
Relative_Compactness -6.790e-01  1.078e-01  -6.299 5.06e-10 ***
Surface_Area         -7.620e-01  1.490e-01  -5.115 3.97e-07 ***
Wall_Area             2.629e-01  2.873e-02   9.153  < 2e-16 ***
Overall_Height        7.237e-01  5.862e-02  12.345  < 2e-16 ***
Glazing_Area          2.632e-01  1.074e-02  24.503  < 2e-16 ***
Glazing_Distribution  3.132e-02  1.074e-02   2.916  0.00365 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2906 on 761 degrees of freedom
Multiple R-squared:  0.9162,    Adjusted R-squared:  0.9155
F-statistic:  1387 on 6 and 761 DF,  p-value: < 2.2e-16
```

We utilized stepwise variable selection to find the best model. The model we found contained all predictors except **Orientation** and **Roof Area**.

Findings

After removing variables based on the stepwise function. The reduced model is **sufficient** due to p-value > 0.05.



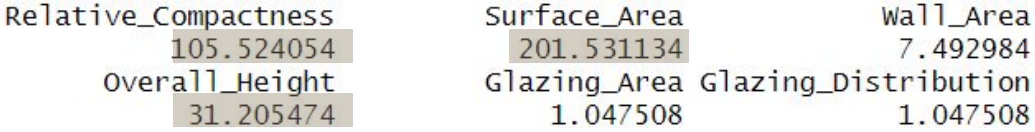
```
Analysis of Variance Table

Model 1: Heating_Load ~ Relative_Compactness + Surface_Area + Wall_Area +
    Roof_Area + Overall_Height + Orientation + Glazing_Area +
    Glazing_Distribution
Model 2: Heating_Load ~ Relative_Compactness + Surface_Area + Wall_Area +
    Overall_Height + Glazing_Area + Glazing_Distribution
  Res.Df    RSS Df  Sum of Sq      F Pr(>F)
1    760 64.273
2    761 64.278 -1 -0.0051323 0.0607 0.8055
```

Reduced

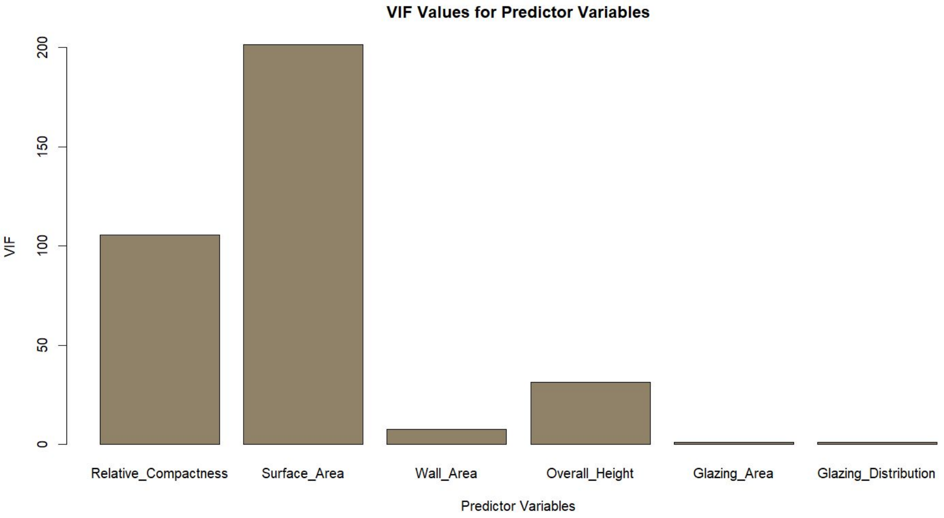
VIF

Half of the current variables pose a problem of multicollinearity (↑ VIF)

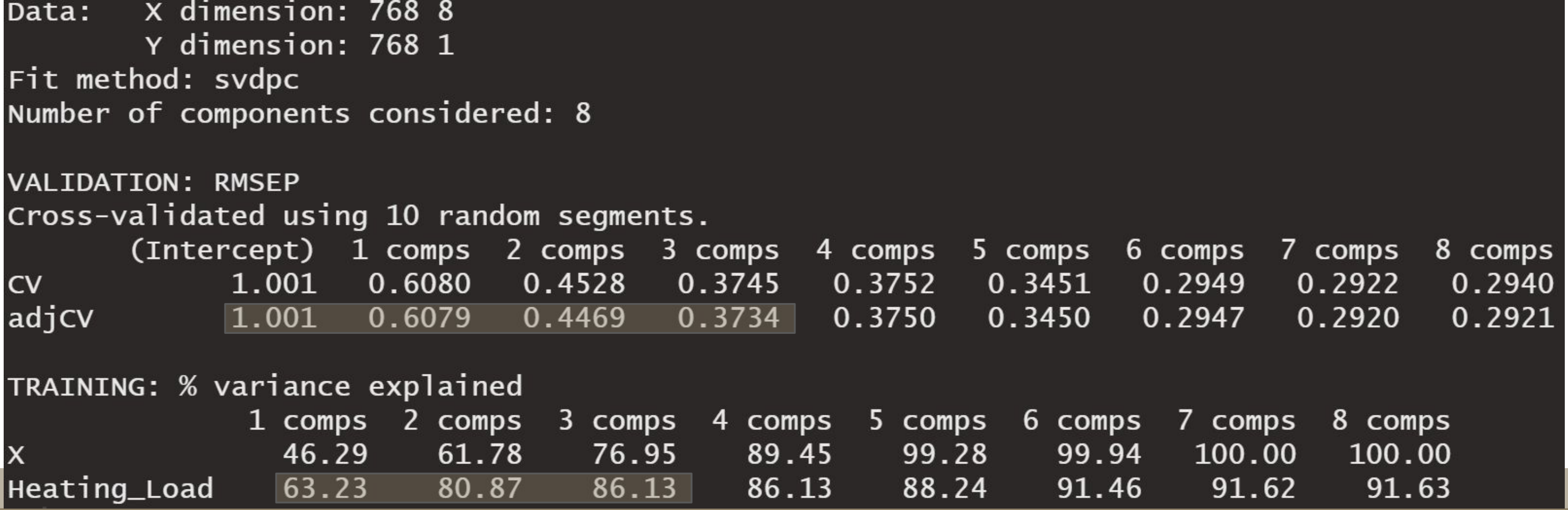
Since we have issues with multicollinearity, we decided to implement **PCR** to mediate these issues.



```
Relative_Compactness          Surface_Area             Wall_Area
          105.524054            201.531134              7.492984
      Overall_Height          Glazing_Area Glazing_Distribution
           31.205474              1.047508              1.047508
```



VIF Values for Predictor Variables



```
Data:   X dimension: 768 8
        Y dimension: 768 1
Fit method: svdpc
Number of components considered: 8

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
CV           1.001   0.6080   0.4528   0.3745   0.3752   0.3451   0.2949   0.2922   0.2940
adjCV        1.001   0.6079   0.4469   0.3734   0.3750   0.3450   0.2947   0.2920   0.2921

TRAINING: % variance explained
              1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X               46.29    61.78    76.95    89.45    99.28    99.94   100.00   100.00
Heating_Load    63.23    80.87    86.13    86.13    88.24    91.46    91.62    91.63
```

Principal Component Regression (PCR)

After creating a PCR model, we must decide on which principle components are valuable. We found that the first three components are sufficient to use. We chose this based on the RMSEP value and Heating Load’s % of variance explained.

Principal Component Regression (PCR)

Now that we have our model and the optimal number of principal components, we can find our coefficient values for each predictor.

From the visual, we can see that **Wall Area**, **Roof Area**, and **Overall Height** are our most influential predictors on our target variable.

After finding our final model, we must test the predictive power. We split the data up into training and test data. We Then made a prediction vector and tested it using the test data. We had an RMSE value of **0.587** which is considerable low.

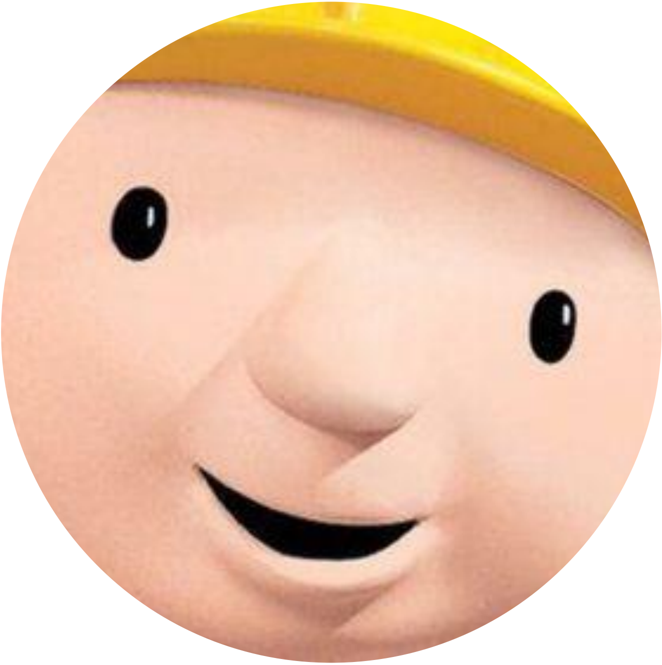
Relative Compactness:	0.112
Surface Area:	-0.119
Wall Area:	0.346
Roof Area:	-0.284
Overall Height:	0.284
Orientation:	-0.021
Glazing Area:	0.140
Glazing Distribution:	0.164

Problems

- Multicollinearity
- Simulated Data
  - lacks randomness

Future Work

- Implement the project using multiple real world data sets instead of a simulated one.
- Experiment with other models



THANKS

Does anyone have any questions?